Published in final edited form as:

J Stat Plan Inference. 2012 October ; 142(10): 2819–2831. doi:10.1016/j.jspi.2012.04.006.

Marginal methods for clustered longitudinal binary data with incomplete covariates

Baojiang Chen^{a,*}, Grace Y. Yi^b, Richard J. Cook^b, and Xiao-Hua Zhou^{c,d}

^aDepartment of Biostatistics, University of Nebraska Medical Center, Omaha, NE 68198, USA

^bDepartment of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

^cDepartment of Biostatistics, University of Washington, Seattle, WA 98195, USA

^dBiostatistics Unit, HSR&D Center of Excellence, VA Puget Sound Health Care System Seattle, WA 98101, USA

Abstract

Many analyses for incomplete longitudinal data are directed to examining the impact of covariates on the marginal mean responses. We consider the setting in which longitudinal responses are collected from individuals nested within clusters. We discuss methods for assessing covariate effects on the mean and association parameters when covariates are incompletely observed. Weighted first and second order estimating equations are constructed to obtain consistent estimates of mean and association parameters when covariates are missing at random. Empirical studies demonstrate that estimators from the proposed method have negligible finite sample biases in moderate samples. An application to the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS) demonstrates the utility of the proposed method.

Keywords

Association; Generalized estimating equation; Longitudinal data; Missing covariates

1. Introduction

In many situations, longitudinal data arise in clusters. Common examples include longitudinal community intervention studies (e.g., Perry et al., 1989), family studies involving repeated assessments of individual members over time (Payment et al., 1991), longitudinal school-based studies in which individual schools are randomized to receive an experimental or control intervention (Cameron et al., 1999), and longitudinal center-based studies for an Alzheimer's disease. Clustered longitudinal data feature both a cross-sectional and a longitudinal correlation structure, and interest often resides in the strength of both types of association. When the association parameters are of central importance, second-order generalized estimating equation (GEE) is needed to perform estimation. With longitudinal binary data, Prentice (1988) developed such equations and emphasized estimation of correlation parameters. Fitzmaurice et al. (1993) proposed a model that parameterizes the association in terms of conditional odds ratios. Lipsitz et al. (1991), Liang et al. (1992), Carey et al. (1993), Molenberghs and Lesaffre (1994), Lang and Agresti

(1994), and Fitzmaurice and Lipsitz (1995) have proposed models that parameterize the association in terms of marginal odds ratios.

In clinical trials and observational studies, however, complete data are often not available for every subject. Missing data may arise in many circumstances, including the unavailability of covariate measurements, survey nonresponse, study subjects failing to report to a clinic for regular evaluations, respondents refusing to answer certain items on a questionnaire, and loss of data. Problems often arise if the missing data mechanism is not missing completely at random (MCAR) (Little and Rubin, 2002). In particular, easily implemented complete case analysis often yields invalid inference. The GEE approach (Liang and Zeger, 1986), on the other hand, may yield consistent estimates for the regression parameters if data are MCAR. When data are missing at random (MAR) or missing not at random (MNAR) (Little and Rubin, 2002), analysis based on GEE may give inconsistent estimators. Robins et al. (1994, 1995) developed a class of estimators based on inverse probability weighted generalized estimating equations (IPWGEEs) under regression settings where data are MAR. This approach involves modeling the missing data process, and requires models for the marginal mean response and missing data process to be correctly formulated. When either one of these models is misspecified, the resultant estimator may be inconsistent. For this situation, doubly robust estimating equations have considerable appeal (e.g., Van der Laan and Robins, 2003; Chen and Zhou, 2011).

Yi and Cook (2002) discussed marginal methods for incomplete responses in longitudinal data arising in clusters, where inverse probability weighted second-order estimating equations were developed for the case of complete covariates. Under a MAR mechanism, this method facilitates consistent estimation of marginal mean parameters and association parameters. Little attention, however, has been directed to addressing the impact of missing covariates on the association parameters in the context of studies with clustered longitudinal data. This paper handles this problem. Weighted first and second order estimating equations are constructed to conduct consistent estimation for both marginal mean and association parameters.

The paper is organized as follows. We introduce basic notation and model formulations in Section 2. In Section 3, we provide the details on estimation and inference procedures for a missing covariate. Numerical studies concerning issues of bias are given in Section 4. Data arising from an Alzheimer's disease study are analyzed with the proposed method in Section 5. Extensions to deal with multiple missing covariates are outlined in Section 6. Concluding remarks are made in Section 7.

2. Notation and model formulation

2.1. Response process

Suppose there are n clusters and J_i individuals within cluster i , $i=1, 2, \dots, n$. We consider the case that all subjects are scheduled to be measured at K time points. The proposed method can be readily modified to accommodating settings with variable number of measurements among subjects. Let $Y_{ij} = (Y_{ij1}, Y_{ij2}, \dots, Y_{ijK})'$ denote the $K \times 1$ binary response vector for subject j in cluster i which we assume that it is always observed. Let $X_{ij} = (X_{ij1}, X_{ij2}, \dots, X_{ijK})'$ be the covariate vector that may be incompletely observed, and $Z_{ijk} = (1, Z_{ijk1}, Z_{ijk2}, \dots, Z_{ijk,p-2})'$ be the covariate vector that is always observed. Denote

$Y_i = (Y'_{i1}, \dots, Y'_{iJ_i})'$, $X_i = (X'_{i1}, \dots, X'_{iJ_i})'$, $Z_{ij} = (Z'_{ij1}, \dots, Z'_{ijK})'$, and $Z_i = (Z'_{i1}, \dots, Z'_{iJ_i})'$. Let $\mu_{ijk} = E(Y_{ijk} | X_{ij}, Z_{ij}) = P(Y_{ijk} = 1 | X_{ij}, Z_{ij})$, $\mu_{ij} = (\mu_{ij1}, \mu_{ij2}, \dots, \mu_{ijK})'$, $j = 1, \dots, J_i$, $i = 1, \dots, n$, and

$\mu_i = (\mu'_{i1}, \dots, \mu'_{iJ_i})'$. We suppose the mean structure of Y_{ijk} depends on the covariate vector for

subject j at time k (e.g., Pepe and Anderson 1994; Robins et al., 1999) and consider a regression model for the mean of the form

$$g(\mu_{ijk}) = X_{ijk}\beta_x + Z'_{ijk}\beta_z,$$

for $k = 1, \dots, K, j = 1, \dots, J_k$, where g is a monotone link function. Here we suppose only one covariate X_{ijk} is potentially missing, but extensions to dealing with both missing responses and multiple missing covariates are given in Section 6. Let $\beta = (\beta_x, \beta_z)'$ be the vector of regression parameters. The variance for the response Y_{ijk} is specified as

$$v_{ijk} = \text{Var}(Y_{ijk}|X_i, Z_i) = \mu_{ijk}(1 - \mu_{ijk}),$$

which depends on the regression parameter vector β .

The joint probability for any pair $(Y_{ijk}, Y_{ij'k'})$ of binary responses, given by

$$\mu_{i,jk;j'k'} = E(Y_{ijk}Y_{ij'k'}|X_i, Z_i) = P(Y_{ijk}=1, Y_{ij'k'}=1|X_i, Z_i), \quad (j, k) \neq (j', k'),$$

can be modeled in terms of the marginal probabilities $\mu_{ijk}(\beta)$ and $\mu_{ij'k'}(\beta)$ and a vector of association parameters. We use odds ratios to characterize the association among responses. The (conditional) odds ratio, $\psi_{i,jk;j'k'}$, between Y_{ijk} and $Y_{ij'k'}$, is defined as

$$\psi_{i,jk;j'k'} = \frac{P(Y_{ijk}=1, Y_{ij'k'}=1|X_i, Z_i)P(Y_{ijk}=0, Y_{ij'k'}=0|X_i, Z_i)}{P(Y_{ijk}=1, Y_{ij'k'}=0|X_i, Z_i)P(Y_{ijk}=0, Y_{ij'k'}=1|X_i, Z_i)}. \quad (1)$$

Regression models for the association are typically specified as

$$\log \psi_{i,jk;j'k'} = u'_{i,jk;j'k'} \cdot \phi,$$

where $u_{i,jk;j'k'}$ is a vector of covariates which specifies the form of the association between Y_{ijk} and $Y_{ij'k'}$, and ϕ is a vector of regression parameters. For example, letting $u_{i,jk;j'k'}$ be the scalar 1 leads to the exchangeable association between responses (Yi and Cook, 2002).

The joint probability $\mu_{i,jk;j'k'}$ is determined by the marginal means and the odds ratio. That is,

$$\mu_{i,jk;j'k'} = \begin{cases} \frac{a_{i,jk;j'k'} - [a_{i,jk;j'k'}^2 - 4\psi_{i,jk;j'k'}(\psi_{i,jk;j'k'} - 1)\mu_{ijk}\mu_{ij'k'}]^{1/2}}{2(\psi_{i,jk;j'k'} - 1)}, & \psi_{i,jk;j'k'} \neq 1, \\ \mu_{ijk} \cdot \mu_{ij'k'}, & \psi_{i,jk;j'k'} = 1, \end{cases}$$

where $a_{i,jk;j'k'} = 1 - (1 - \psi_{i,jk;j'k'}) (\mu_{ijk} + \mu_{ij'k'})$ (e.g. Lipsitz et al., 1991).

2.2. Missing data models

Let $R_{ij} = (R_{ij1}, R_{ij2}, \dots, R_{ijK})'$ be the vector of missing data indicators, defined as $R_{ijk} = 1$ if X_{ijk} is observed and 0 otherwise. Let $r_{ij} = (r_{ij1}, \dots, r_{ijK})'$ be a realization of R_{ij} and let

$H_{ijk}^r = \{r_{ij1}, \dots, r_{ij,k-1}\}$ denote the history of the missing data indicators for subject j up to but not including visit k , $k=2, 3, \dots, K$. For ease of presentation, with $k=1$ let H_{ijk}^r be a void set, $j = 1, \dots, J_i$.

We focus on dealing with the MAR mechanism in marginal models, where we assume that,

$$P(R_{ij}=r_{ij}, R_{ij'}=r_{ij'} | Y_i, X_i, Z_i) = P(R_{ij}=r_{ij}, R_{ij'}=r_{ij'} | Y_i, X_i^{(o)}, Z_i). \quad (2)$$

Here $X_i^{(o)}$ represents the observed components of X_i . A conditional approach is employed to model the missing data mechanism in (2). Namely, we let

$$P(R_{ij}=r_{ij} | Y_i, X_i, Z_i) = \prod_{k=1}^K P(R_{ijk}=r_{ijk} | H_{ijk}^r, Y_i, X_i, Z_i),$$

and

$$P(R_{ij}=r_{ij}, R_{ij'}=r_{ij'} | Y_i, X_i, Z_i) = \prod_{k=1}^K P(R_{ijk}=r_{ijk}, R_{ij'k}=r_{ij'k} | H_{ijk}^r, H_{ij'k}^r, Y_i, X_i, Z_i).$$

Instead of modeling the joint probability $P(R_{ij}=r_{ij}, R_{ij'}=r_{ij'} | Y_i, X_i, Z_i)$ for R_{ij} and $R_{ij'}$ over the entire course, we introduce parametric models for the joint conditional probability

$P(R_{ijk}=r_{ijk}, R_{ij'k}=r_{ij'k} | H_{ijk}^r, H_{ij'k}^r, Y_i, X_i, Z_i)$ at each observation time which we assume

$$P(R_{ijk}=r_{ijk}, R_{ij'k}=r_{ij'k} | H_{ijk}^r, H_{ij'k}^r, Y_i, X_i, Z_i) = P(R_{ijk}=r_{ijk}, R_{ij'k}=r_{ij'k} | H_{ijk}^r, H_{ij'k}^r, Y_i, X_i^{(o)}, Z_i), \quad (3)$$

for for each k . Note that (3) implies (2), but not vice versa. Although mechanism (2) covers a larger class of MAR models than (3), many realistic and practical models can be embedded in the class characterized by (3), and this strategy has been commonly used to deal with MAR missing data processes (see Robins et al., 1995). Finally, we note that for MAR mechanisms, it is often convenient to adopt the further assumption that the missing data indicators at time k depend only on the previously observed outcomes and covariates, and we do this in Sections 4 and 5.

For subject j in cluster i and time k , let $\lambda_{ijk} = P(R_{ijk}=1 | H_{ijk}^r, Y_i, X_i, Z_i)$ be the conditional probability, given the history of the indicator variable and entire vectors of responses and covariates. A logistic link may be used to give

$$\text{logit}(\lambda_{ijk}) = u_{ijk}^* \cdot \gamma,$$

where u_{ijk}^* reflects relevant information in $\{Y_i, X_i^{(o)}, Z_i, H_{ijk}^r\}$.

We assume that

$$P(R_{ijk}=1 | H_{ijk}^r, H_{ij'k}^r, Y_i, X_i, Z_i) = P(R_{ijk}=1 | H_{ijk}^r, Y_i, X_i, Z_i) \quad \text{for } j \neq j', \quad (4)$$

for each time point k . To incorporate possible dependence between paired missing data indicators at a give time point, we define the odds ratio as

$$\psi_{ijk;j'k}^* = \frac{P(R_{ijk}=1, R_{ij'k}=1|L_{ijk;j'k}) \cdot P(R_{ijk}=0, R_{ij'k}=0|L_{ijk;j'k})}{P(R_{ijk}=1, R_{ij'k}=0|L_{ijk;j'k}) \cdot P(R_{ijk}=0, R_{ij'k}=1|L_{ijk;j'k})},$$

for $k = 1, \dots, K$, where $L_{ijk;j'k} = \{H_{ijk}^r, H_{ij'k}^r, Y_i, X_i, Z_i\}$. Regression models may be used to characterize the effect of covariates on the odds ratio $\psi_{ijk;j'k}^*$ for each fixed time point k . Let ϕ^* be the regression parameters linking the odds ratios $\psi_{ijk;j'k}^*$ to the associated covariate vector $u_{ijk;j'k}^*$, say, through the log link, for example,

$$\log(\psi_{ijk;j'k}^*) = u_{ijk;j'k}^{*'} \cdot \phi^*.$$

Then $\alpha = (\gamma', \phi^{*'})'$ is the vector of parameters associated with the missing data process.

Let $\lambda_{ijk;j'k} = P(R_{ijk}=1, R_{ij'k}=1|H_{ijk}^r, H_{ij'k}^r, Y_i, Z_i, X_i)$ be the joint conditional probability for $(R_{ijk}, R_{ij'k})$, given by

$$\lambda_{ijk;j'k} = \begin{cases} \frac{a_{ijk;j'k}^* - \left[a_{ijk;j'k}^{*2} - 4\psi_{ijk;j'k}^* (\psi_{ijk;j'k}^* - 1) \lambda_{ijk} \lambda_{ij'k} \right]^{1/2}}{2(\psi_{ijk;j'k}^* - 1)} & \text{for } \psi_{ijk;j'k}^* \neq 1, \lambda_{ijk} \cdot \lambda_{ij'k} \\ \psi_{ijk;j'k}^* & \text{for } \psi_{ijk;j'k}^* = 1, \end{cases}$$

where $a_{ijk;j'k}^* = 1 - (1 - \psi_{ijk;j'k}^*)(\lambda_{ijk} + \lambda_{ij'k})$.

For cluster i , let $\pi_{ijk} = P(R_{ijk} = 1|Y_i, X_i, Z_i)$ be the marginal probability X_{ijk} is observed, and $\pi_{ijk;j'k} = P(R_{ijk} = 1, R_{ij'k} = 1|Y_i, X_i, Z_i)$ be the joint probability that X_{ijk} and $X_{ij'k}$ are observed. These marginal probabilities can be expressed in terms of the conditional probabilities λ_{ijk} and $\lambda_{ijk;j'k}$. That is, the marginal probabilities π_{ijk} ($k = 2$) can be written as

$$\pi_{ijk} = \sum_{H_{ijk}^r} \left[\lambda_{ijk} \cdot \prod_{l=1}^{k-1} \left[\lambda_{ijl}^{r_{ijl}} (1 - \lambda_{ijl})^{1-r_{ijl}} \right] \right].$$

For $j = j'$ and $k < K$, the joint probability $\pi_{ijk;j'k}$ is written as

$$\begin{aligned} \pi_{ijk;j'k} &= \sum_{r_{ij,k+1} \dots r_{ij,k'-1}, H_{ijk}^r} \left[P(R_{ijk'}=1|r_{ij,k+1} \dots r_{ij,k'-1}, R_{ijk}=1, H_{ijk}^r, Y_i, X_i, Z_i) \right. \\ &\quad \left. \dots P(R_{ij,k+1}=r_{ij,k+1}|R_{ijk}=1, H_{ijk}^r, Y_i, X_i, Z_i) \cdot P(R_{ijk}=1|H_{ijk}^r, Y_i, X_i, Z_i) \cdot P(H_{ijk}^r|Y_i, X_i, Z_i) \right] \\ &= \sum_{r_{ij,k+1} \dots r_{ij,k'-1}, H_{ijk}^r} \left[\tilde{\lambda}_{ijk'} \cdot \prod_{l=k+1}^{k'-1} \tilde{\lambda}_{ijl}^{r_{ijl}} (1 - \tilde{\lambda}_{ijl})^{1-r_{ijl}} \cdot \lambda_{ijk} \cdot \prod_{l=1}^{k-1} \lambda_{ijl}^{r_{ijl}} (1 - \lambda_{ijl})^{1-r_{ijl}} \right], \end{aligned}$$

where $\tilde{\lambda}_{ijk'}$ denotes $\lambda_{ijk'}$ by conditioning that $R_{ijk} = 1$ for $k' > k$.

For $j < j'$ and $k = k'$, we have

$$\begin{aligned}\pi_{i;jk;j'k} &= \sum_{H_{ijk}^r, H_{ij'k}^r} \left\{ P(R_{ijk}=1, R_{ij'k}=1 | H_{ijk}^r, H_{ij'k}^r, Y_i, X_i, Z_i) \cdot \prod_{l=1}^{k-1} P(R_{ijl}=r_{ijl}, R_{ij'l}=r_{ij'l} | H_{ijl}^r, H_{ij'l}^r, Y_i, X_i, Z_i) \right\} \\ &= \sum_{H_{ijk}^r, H_{ij'k}^r} \left\{ \lambda_{i;jk;j'k} \cdot \prod_{l=1}^{k-1} \left[\lambda_{i;jl;j'l}^{r_{ijl} r_{ij'l}} \cdot (\lambda_{ijl} - \lambda_{i;jl;j'l})^{r_{ijl}(1-r_{ij'l})} \cdot (\lambda_{ij'l} - \lambda_{i;jl;j'l})^{r_{ij'l}(1-r_{ijl})} \cdot (1 - \lambda_{ijl} - \lambda_{ij'l} + \lambda_{i;jl;j'l})^{(1-r_{ijl})(1-r_{ij'l})} \right] \right\}.\end{aligned}$$

For $j < j'$ and $k < k'$, we have

$$\begin{aligned}\pi_{i;jk;k'} &= \sum_{H_{ijk}^r, H_{ij'k'}^r} P(R_{ij'k'}=1, R_{ijk}=1, H_{ijk}^r, H_{ij'k'}^r | Y_i, X_i, Z_i) = \sum_{H_{ijk}^r, H_{ij'k'}^r} \left\{ P(R_{ij'k'}=1 | H_{ij'k'}^r, R_{ijk}=1, H_{ijk}^r, Y_i, X_i, Z_i) \right. \\ &\quad \cdot \left[\prod_{l=k+1}^{k'-1} P(R_{ij'l}=r_{ij'l} | H_{ij'l}^r, R_{ijk}=1, H_{ijk}^r, Y_i, X_i, Z_i) \right] \cdot P(R_{ijk}=1, R_{ij'k}=r_{ij'k} | H_{ijk}^r, H_{ij'k}^r, Y_i, X_i, Z_i) \\ &\quad \cdot \prod_{l=1}^{k-1} P(R_{ijk}=r_{ijl}, R_{ij'l}=r_{ij'l} | H_{ijl}^r, H_{ij'l}^r, Y_i, X_i, Z_i) \left. \right\} \\ &= \sum_{H_{ijk}^r, H_{ij'k'}^r} \left\{ \lambda_{ij'k'} \cdot \left[\prod_{l=k+1}^{k'-1} \lambda_{ij'l}^{r_{ij'l}} (1 - \lambda_{ij'l})^{1-r_{ij'l}} \right] \cdot \lambda_{i;jk;j'k}^{r_{ij'k}} (\lambda_{ijk} - \lambda_{i;jk;j'k})^{1-r_{ij'k}} \right. \\ &\quad \cdot \prod_{l=1}^{k-1} \left[\lambda_{i;jl;j'l}^{r_{ijl} r_{ij'l}} \cdot (\lambda_{ijl} - \lambda_{i;jl;j'l})^{r_{ijl}(1-r_{ij'l})} \cdot (\lambda_{ij'l} - \lambda_{i;jl;j'l})^{r_{ij'l}(1-r_{ijl})} \cdot (1 - \lambda_{ijl} - \lambda_{ij'l} + \lambda_{i;jl;j'l})^{(1-r_{ijl})(1-r_{ij'l})} \right] \left. \right\};\end{aligned}$$

the third equality holds because of assumption (4), where we have

$$P(R_{ij'k'}=1 | H_{ij'k'}^r, R_{ijk}=1, H_{ijk}^r, Y_i, X_i, Z_i) = P(R_{ij'k'}=1 | H_{ij'k'}^r, Y_i, X_i, Z_i).$$

Similarly, we can write the corresponding formula for $j < j'$ and $k > k'$.

3. Estimation and inference

Our primary interest lies in estimating the parameter β of the marginal mean regression model as well as the association parameter ϕ . Let $\theta = (\beta', \phi')'$.

3.1. Estimating equations for the response parameters

Here we develop an estimation method by including a weight to the usual GEE to adjust for the effects of missingness of the covariate variable. If

$$\Delta_i(\alpha) = \text{diag} \left\{ I(R_{ijk}=1) / \pi_{ijk}, 1 \leq j \leq J_i, 1 \leq k \leq K \right\}$$

is a $J_i K \times J_i K$ weight matrix, then the weighted GEEs for β are given by

$$U_1(\theta, \alpha) = \sum_{i=1}^n U_{1i}(\theta, \alpha) = 0, \quad (5)$$

where $U_{1i}(\theta, \alpha) = D_i V_i^{-1} \Delta_i(\alpha) (Y_i - \mu_i)$, $D_i = \partial \mu_i' / \partial \beta$ be the $p \times J_i K$ derivative matrix of the mean vector μ_i with respect to β , and $V_i = \text{diag}(\mu_{ijk}(1-\mu_{ijk}), 1 \leq j \leq J_i, 1 \leq k \leq K)$.

For the association parameters ϕ , let

$$\Delta_i^*(\alpha) = \text{diag} \left\{ I(R_{ijk}=1, R_{ij'k'}=1) / \pi_{ijk;j'k'}, (j, k) < (j', k') \right\},$$

be the weight matrix, where we define $(j, k) < (j', k')$ if $j < j'$, or $j = j'$ and $k < k'$. The weighted GEEs for ϕ are of the form

$$U_2(\theta, \alpha) = \sum_{i=1}^n U_{2i}(\theta, \alpha) = 0, \quad (6)$$

where

$U_{2i}(\theta, \alpha) = C_i W_i^{-1} \cdot \Delta_i^*(\alpha) \cdot (\mathcal{Y}_i - \xi_i)$, $\mathcal{Y}_i = (Y_{ijk} Y_{ij'k'}, (j, k) < (j', k'))'$, $\xi_i = E(\mathcal{Y}_i | X_i, Z_i) = (\mu_{ijk;j'k'}, (j, k) < (j', k'))'$, $C_i = \partial \xi_i / \partial \phi$ is the derivative matrix of the mean vector ξ_i with respect to ϕ , and W_i is a working covariance matrix. The true covariance matrix of $\mathcal{Y}_i$ involves third and fourth moments of the responses, which we would rather not estimate. With binary responses, the working independence covariance matrix

$$W_i = \text{diag} \left\{ \mu_{ijk;j'k'} \cdot (1 - \mu_{ijk;j'k'}), (j, k) < (j', k') \right\}$$

is often specified.

3.2. Estimation for parameters of the missing-data process

Let $R_{i:k} = (R_{i1k}, \dots, R_{iJ_k k})'$ indicate the missing pattern for J_i subjects in cluster i at time k , and $R_i = (R'_{i1}, \dots, R'_{iK})'$. Let $\Lambda_{i:k} = (\lambda_{i1k}, \dots, \lambda_{iJ_k k})'$ be the vector of conditional expectations of $R_{i:k}$, and $\Lambda_i = (\Lambda'_{i1}, \dots, \Lambda'_{iK})'$. Because $R_{i:k}$ is a vector of binary variables, we specify its covariance matrix as $V_{i:k} = (v_{ijk;j'k'}^*)$, where $v_{ijk;j'k'}^* = \lambda_{ijk}(1 - \lambda_{ijk})$ if $j = j'$ and $\lambda_{ijk;j'k'} - \lambda_{ijk} \cdot \lambda_{ij'k}$ otherwise. Let $V_i^* = \text{diag}(V_{i:k}, k=1, \dots, K)$ be the covariance matrix for R_i . Define $D_i^* = \partial \Lambda_i' / \partial \gamma$, then the estimating equations for γ are of the following form:

$$\sum_{i=1}^n S_{1i}(\alpha) = 0, \quad (7)$$

where $S_{1i}(\alpha) = D_i^* V_i^{*-1} (R_i - \Lambda_i)$.

For the second order estimating equations for the association parameter ϕ^* , we define

$\mathcal{R}_{i:k} = (R_{i1k} R_{i2k}, \dots, R_{iJ_{i-1,k} k} R_{iJ_k k})'$, and $\mathcal{R}_i = (\mathcal{R}_{i1}, \dots, \mathcal{R}_{iK})'$. Let $\Lambda_i^* = (\Lambda_{i1}^*, \dots, \Lambda_{iK}^*)'$, where $\Lambda_{i,k}^* = (\lambda_{ijk;j'k'}, j < j')'$, $C_i^* = \partial \Lambda_i^* / \partial \phi^*$, and

$$W_i^* = \text{diag}(\Lambda_i \cdot (1 - \Lambda_i)).$$

Then estimating equations for ϕ^* are given by

$$\sum_{i=1}^n S_{2i}(\alpha) = 0, \quad (8)$$

where $S_{2i}(\alpha) = C_i^* W_i^{*-1} \cdot (\mathcal{R}_i - \Lambda_i^*)$.

3.3. Estimation and inference

We employ a Fisher-scoring algorithm for estimation of $\vartheta = (\alpha', \theta')'$. To do this, let

$$H_i(\vartheta) = \begin{pmatrix} S_i(\alpha) \\ U_i(\theta, \alpha) \end{pmatrix}, \quad M^*(\alpha) = \sum_{i=1}^n \partial S_i(\alpha) / \partial \alpha',$$

and $M(\vartheta) = \sum_{i=1}^n \partial U_i(\vartheta) / \partial \theta'$, where $S_i(\alpha) = (S'_{1i}(\alpha), S'_{2i}(\alpha))'$ and

$U_i(\theta, \alpha) = (U'_{1i}(\theta, \alpha), U'_{2i}(\theta, \alpha))'$. As the estimating functions $S_i(\alpha)$ are free of θ , the derivative matrix $H_i(\vartheta) / \vartheta$ is lower triangular, i.e.,

$$\frac{\partial H_i(\vartheta)}{\partial \vartheta'} = \begin{pmatrix} \partial S_i(\alpha) / \partial \alpha' & 0 \\ \partial U_i(\theta, \alpha) / \partial \alpha' & \partial U_i(\theta, \alpha) / \partial \theta' \end{pmatrix}.$$

Given an initial value $\vartheta^{(0)}$, updated estimates of ϑ are obtained with the following iterative equation:

$$\vartheta^{(t+1)} = \vartheta^{(t)} - \begin{pmatrix} M^*(\alpha^{(t)}) & 0 \\ \sum_{i=1}^n [\partial U_i(\vartheta) / \partial \alpha']|_{\vartheta^{(t)}} & M(\vartheta^{(t)}) \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n S_i(\alpha^{(t)}) \\ \sum_{i=1}^n U_i(\vartheta^{(t)}) \end{pmatrix}, \quad (9)$$

$t = 0, 1, \dots$, where $\widehat{\vartheta}$ denotes the value upon convergence.

Alternatively, one can invoke a two-stage estimation procedure in the same spirit of Yi and Cook (2002). Under this scheme an estimate of α is obtained as the solution to

$\sum_{i=1}^n S_i(\alpha) = 0$, and then to solve $\sum_{i=1}^n U_i(\theta, \widehat{\alpha}) = 0$ where $\widehat{\alpha}$ is used in place of α in (5) and (6). This two-stage estimation procedure can be realized using iterative equations:

$$\alpha^{(s+1)} = \alpha^{(s)} - [M^*(\alpha^{(s)})]^{-1} \cdot \sum_{i=1}^n S_i(\alpha^{(s)}), \quad s = 0, 1, \dots \quad (10)$$

and

$$\theta^{(t+1)} = \theta^{(t)} - [M(\theta^{(t)}, \widehat{\alpha})]^{-1} \cdot \sum_{i=1}^n U_i(\theta^{(t)}, \widehat{\alpha}), \quad t = 0, 1, \dots \quad (11)$$

The two-stage iterative equations (10) and (11) differ from joint iterative equation (9). Even under the special situation that the components of the left lower corner are zero in the inverse matrix of (9), (9) does not necessarily yield the same updated values as those from (10) and (11). However, the updated values from these two procedures converge to the same limit under certain regularity conditions.

While the two-stage procedure based on (10) and (11) is much easier to use for estimation of $\widehat{\theta}$, the joint formulation based on $H(\theta)$ is more useful for developing the asymptotic distribution for $\widehat{\theta}$. Since $E[H(\theta)] = 0$, by Theorem 3.4 of Newey and Mcfadden (1994), under standard regularity conditions, there is a unique solution $\widehat{\theta}$ to the equation

$\sum_{i=1}^n H_i(\theta) = 0$ with probability approaching 1, which satisfies

$$n^{1/2}(\widehat{\theta} - \theta) = - \left\{ E \left[\partial H_i(\theta) / \partial \theta' \right] \right\}^{-1} \cdot n^{-1/2} \sum_{i=1}^n H_i(\theta) + o_p(1).$$

For the estimator $\widehat{\theta}$, we have

$$n^{1/2}(\widehat{\theta} - \theta) = - \Gamma^{-1} n^{-1/2} \cdot \sum_{i=1}^n Q_i(\theta, \alpha) + o_p(1),$$

where $\Gamma = E[U(\theta, \alpha)U(\theta, \alpha)']$, and $Q(\theta, \alpha) = U(\theta, \alpha) - E[U(\theta, \alpha) | g = a] \cdot [E[S(\alpha) | g = a]]^{-1} \cdot S(\alpha)$. The central limit theorem then leads to the asymptotic distribution for $n^{1/2}(\widehat{\theta} - \theta)$, which is normal with mean 0 and asymptotic variance $\Gamma^{-1} \Sigma \Gamma^{-1}$, where $\Sigma = E[Q(\theta, \alpha)Q(\theta, \alpha)']$.

4. Empirical studies

4.1. Simulation studies

In the simulation study, we focus on a drop-out setting with $K=3$ and $J=3$, $i=1, 2, \dots, n$, and $n=500$. We simulate longitudinal binary responses from the model with

$$\text{logit}(\mu_{ijk}) = \beta_0 + \beta_1 x_{ijk1} + \beta_2 x_{ijk2} + \beta_3 x_{ijk3}, \quad (12)$$

where x_{ijk1} is a time varying binary covariate subject to missingness, independently generated from $\text{Bin}(1, 0.5)$, $x_{ijk2} = I(k=2)$, and $x_{ijk3} = I(k=3)$. Take $\text{expit}(\beta_0) = 0.4$, $\text{expit}(\beta_0 + \beta_2) = 0.5$, $\text{expit}(\beta_0 + \beta_3) = 0.6$, and $\text{exp}(\beta_1) = 0.5$, where $\text{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. The association between responses is specified through the odds ratios given by (1) as $\Psi_{i;jk;j'k'} = 1.2$ if $j \neq j', k = k'$, $\Psi_{i;jk;j'k'} = 1.5$ if $j = j'$, and $\Psi_{i;jk;j'k'} = 2.0$ if $k \neq k'$.

We generate missing indicators by the following model:

$$\text{logit}(\lambda_{ijk}) = \alpha_0 + \alpha_1 y_{ij,k-1} + \alpha_2 y_{ijk} + \alpha_3 x_{ij,k-1,1}, \quad k=2, 3, \quad (13)$$

where we assume the data are always observed at the first time point. The true values are taken as $\text{expit}(\alpha_0) = 0.7$, and $\text{exp}(\alpha_1) = 0.75$. Here α_2 and α_3 are used to adjust for the missingness proportion. For the associations, we assume the odds ratio at each time point is $\psi_{i;jk;j'k'}^*$, $k=2, 3$, and vary it to indicate cross-sectional dependence in the missing data process. The simulation and data analyses are implemented using the statistical soft package R. The code is available from the first author upon request.

We compare two methods in what follows. The first method, using “independence weights”, is the method in which we ignore the cross-sectional associations in the missing-data process, i.e. we set $\psi_{i;jk;j'k'}^* = 1$. The second one, using “clustered weights”, is the proposed method, i.e. the method based on the second order estimating equations accommodating

cross-sectional association within clusters. Tables 1 and 2 report the results, where Bias% is the percent relative bias, ASE is the average standard error based on the sandwich variance estimator, SD is the standard deviation for the 2000 simulations, and CP% is the percent coverage probability for the 95% confidence intervals. It is seen that both methods give negligible empirical biases for the regression coefficients and there is no much difference in associated standard errors, which is partly due to the same equation form for β for these two methods.

However, for the association parameters, when there is cross-sectional association within clusters in the missing data model, the independence weights analysis gives bigger biases for the mixed association parameter $\Psi_{ijk;j'k'}$ and cross-sectional association parameter $\Psi_{ijk;j'k}$ estimates and negligible biases for the longitudinal association parameter $\Psi_{ijk;jk}$ estimate. When there is no cross-sectional association within clusters in the missing data model, i.e.

$\psi_{i,j2;j'2}^* = 1$ and $\psi_{i,j3;j'3}^* = 1$, the independence weights method gives negligible biases. But for the clustered weights analysis, it gives consistent estimates in all the cases we consider.

4.2. Asymptotic bias under model misspecification

In the presence of incomplete covariates, valid inferences about the parameter θ should be based on correct specification of the mean and association structures for the response model as well as for the missing data model. Here we investigate the impact of model misspecification for the missing data process.

Let $\hat{\theta}^\dagger$ denote the estimator for θ when the missing data process is misspecified. To characterize the asymptotic bias of $\hat{\theta}^\dagger$, we use the methods of White (1982) to find the value to which $\hat{\theta}^\dagger$ converges. In the spirit of Rotnitzky and Wypij (1994), Fitzmaurice et al. (1995) and Cook et al. (2004), we take the expectation of $U(\theta, \alpha)$ with respect to the true distribution of $G = (R_i, Y_i, X_i, Z_i)$ and set it equal to zero. The solution to this equation, denoted $\theta^\dagger$, is the value to which $\hat{\theta}^\dagger$ converges in probability. If $\mathcal{G}$ is the sample space for G , and $P(g; \alpha, \theta)$ is the true probability of observing the realized value g of G , then solving the equation

$$\sum_{g \in \mathcal{G}} U_i \cdot P(g; \alpha, \theta) = 0 \quad (14)$$

gives the relationship between θ and $\theta^\dagger$, and enables one to characterize the asymptotic bias.

As Chen et al. (2010a,b) and Chen and Zhou (2011) considered, misspecification of the missing data model normally produces biased estimates for the mean parameters β . Here we focus investigating the impact of model misspecification on estimation of the association parameters. The model setup is similar to that used in the section describing the simulation studies. Generally, we assume the true response model is (12) with association parameters $\Psi_{ijk;j'k'}(j, j', k, k')$, $\Psi_{ijk;j'k}(j, j')$ and $\Psi_{ijk;jk}(k, k')$, which are correctly specified. For the missing data models we assume (13) is the true model and correctly specified but the cross-sectional association parameters $\psi_{ijk;j'k}^*$ are misspecified. We evaluate the asymptotic biases for the association parameters $\Psi_{ijk;j'k'}$ induced by ignoring the association structure in the missing data models (i.e., set $\psi_{ijk;j'k}^* = 1$). We set $\alpha_2 = \log(0.5)$ and $\psi_{ijk;j'k}^* = \psi$ and let Ψ vary from 1 to 10.

Fig. 1 contains plots of the asymptotic percent relative bias of the mixed association parameter $\Psi_{ijk;j'k'}(j, j', k, k')$, defined as $100 \times \left(\frac{\psi_{ijk;j'k'}^\dagger - \psi_{ijk;j'k'}}{\psi_{ijk;j'k'}} \right)$, against ψ . For

each plot, three lines are given corresponding to $\alpha_3 = \log(0.5)$, $\log(1.0)$ and $\log(2.0)$. It can be seen that when the association between the missing data indicators increases or the missing proportion increases (α_3 increases), the asymptotic relative bias increases. Fig. 2 contains plots of the asymptotic percent relative bias of the cross-sectional association parameter $\psi_{ijk;j'k}(j, j')$ against ψ . Similarly, when the association between the missing data indicators increases, the asymptotic relative bias increases. As α_3 increases, the asymptotic relative bias $\psi_{ijk;j'k}$ decreases. We also plot the asymptotic percent relative bias of the longitudinal association parameter $\psi_{ijk;jk'}(k, k')$ against ψ . No apparent bias is observed for this parameter estimate. This is not surprising since we assume the longitudinal association in the missing data model is correctly specified but with the cross-sectional association misspecified. When longitudinal association is misspecified in the missing data process, the induced longitudinal association parameter $\psi_{ijk;jk'}(k, k')$ estimate will be biased.

5. Application to an Alzheimer's disease study

A Uniform Data Set (UDS), collected by National Alzheimer's Coordinating Center (NACC), can be used to investigate risk factors that influence the occurrence of dementia. The response in this context is the diagnosis of dementia (yes/no). The covariates that may influence the status of dementia include sex, congestive heart failure (CVCHF, yes/no), family history of dementia (FHDEM, yes/no), diabetes (yes/no), behavioral assessment (depression or dysphoria, yes/no), hypertension (yes/no), education (years), Mini-Mental State Exam (MMSE) score, and age. There are 9384 subjects from 32 Alzheimer's disease centers included at the entry of this study. Follow-up visits for subjects are scheduled at approximately one-year intervals, with up to four clinical visits until 2008. Although the visits are scheduled at 12 months intervals, in reality patients may come to visit at various times. To deal with this we use the calendar year (discrete) for the visit time, which seems reasonable for slowly progressing diseases such as Alzheimer's disease. We found that a very small proportion (less than 1%) of patients exhibit intermittent visits. These subjects were discarded in the analysis. The behavioral assessment covariate is incompletely observed: about 6.7% of subjects did not complete the behavioral assessment at the first follow-up visit; about 4.8% of subjects missed it at the second follow-up visit; about 3.9% of subjects did not have it at the third follow-up visit.

Let $Y_{ijk} = 1$ if subject j was demented at visit k in center i , and 0 otherwise, $k=1, 2, 3, 4$. Consider the regression model for the response process

$$\text{logit } \mu_{ijk} = X'_{ijk} \beta,$$

where X_{ijk} is the covariate vector for subject j at time point k , which includes the function of sex, CVCHF, FHDEM, diabetes, depression, hypertension, education, MMSE, and age. For the associations, we assume the model

$$\log \psi_{ijk;j'k'} = \phi_0 + \phi_1 \cdot I(j=j') + \phi_2 \cdot I(k=k').$$

For the missingness indicators of the covariate depression, we consider models

$$\text{logit } (\lambda_{ijk}) = Z'_{ijk} \alpha, \quad (15)$$

where Z_{ijk} include function of history of the missing indicators, sex, CVCHF, FHDEM, diabetes, observed depression, hypertension, education, MMSE, and age. For the associations, we assume the model

$$\log(\psi_{ijk:j'k}^*) = \phi_k^*,$$

for $k=2, 3, 4$.

Table 3 records the analyses results for the response models. We conduct three analyses here, corresponding to an unweighted analysis, an analysis based on independence weights and an analysis with weights accommodating a within clustering association. For the response model, the analyses based on independence weights and clustered weights give very similar estimates for the mean parameters (β) but quite different estimates for the association parameters. All estimates, however, are different from the estimates based on the unweighted analysis. All three methods reveal that SEX, CVCHF, family history of dementia and education have no significant impact on dementia; depression has a negative effect on the onset of dementia; MMSE has a positive effect to protect the onset of dementia; diabetes and hypertension have positive effects to protect the onset of dementia; while age has a negative effect on protection the onset of dementia. The estimates of the association parameters for the response model suggests that there is strong longitudinal association (ϕ_2 is significant) but no cross-sectional association (ϕ_1 is not statistically significant). Moreover, there are strong mixed associations (ϕ_0 is significant).

For the missing data model, we carry out standard diagnostic tests for the fit of regression models by comparing a model with an expanded model to do a model selection. Here, we only list the results for the final model reported in Table 4. For the mean models, the independent and clustered weight analyses give very close results. Significance of previous observed depression suggests that a missing at random mechanism is perhaps reasonable for the missing covariate process. The significance of ϕ_2^* , ϕ_3^* and ϕ_4^* reveals that there is cross-sectional association of clustering in the missing data process within centers. Significance of SEX indicates that female students are less likely to participate in the study compared to male students. CVCHF, family history of dementia, hypertension and education have no significant effect on the missingness. Significance of previous observed depression suggests that patients are more likely to remain in the study if they were depressed previously. Significance of MMSE suggests that patients are less likely to participate in the study if they have higher MMSE scores. Diabetes and age have a negative effect for the patients on the response to the visit. Significance of visits suggests that patients are more likely to participate in the study as time goes on.

6. Extension for handling multiple missing covariates

In this section, we extend the proposed method to accommodate circumstances in which multiple covariates could be missing. Slightly different notation and methods are used in this section due to the greater complexity. Without loss of generality, we assume that each component of the vector $X_{ijk} = (X_{ijk1}, \dots, X_{ijkp1})'$ may be incompletely observed at visit k , and the covariate vector $Z_{ijk} = (1, Z_{ijk1}, \dots, Z_{ijkp2})'$ is always observed, $k=1, \dots, K$.

Let R_{ijkI} be the indicator that covariate X_{ijkI} is missing, $I=1, \dots, p1$, and $R_{ijk} = (R_{ijk1}, \dots, R_{ijkp1})'$. Assume $R_{ij1} = \mathbf{1}$, where $\mathbf{1}$ is a $p1 \times 1$ vector with each element 1. Let

$H_{ijk}^r = \{r_{ij1}, \dots, r_{ij,k-1}\}$ be the history of missing indicator vector r_{ijk} at visit k , where we use small letter r_{ijk} to denote a realization of the random vector R_{ijk} .

Let $\pi_{ijk} = P(R_{ijk} = 1 | Y_i, X_i, Z_i)$ be the marginal probability that all data are completely observed at time point k for subject j in cluster i . Under a MAR mechanism with

$P(R_{ijk} = r_{ijk} | H_{ijk}^r, Y_i, X_i, Z_i) = P(R_{ijk} = r_{ijk} | H_{ijk}^r, Y_i, X_i^{(o)}, Z_i)$, we write

$$\pi_{ijk} = \sum_{H_{ijk}^r} \left\{ P(R_{ijk} = 1 | H_{ijk}^r, Y_i, X_i, Z_i) \cdot \prod_{l=1}^{k-1} P(R_{ijl} = r_{ijl} | H_{ijl}^r, Y_i, X_i, Z_i) \right\}, \quad (16)$$

for $k \geq 2$, where $\pi_{ij1} = 1$ is assumed. To determine π_{ijk} , we model the joint probability

$P(R_{ijl} = r_{ijl} | H_{ijl}^r, Y_i, X_i, Z_i)$. More specifically, let $\lambda_{ijkl} = P(R_{ijkl} = 1 | H_{ijkl}^r, Y_i, X_i, Z_i)$ for $l = 1, \dots, p_1$.

Let $R_{ijkl}^* = (R_{ijkl} - \lambda_{ijkl}) / \sqrt{\lambda_{ijkl}(1 - \lambda_{ijkl})}$, $\rho_{ijkst} = E(R_{ijks}^* R_{ijkt}^*)$, and $\rho_{ijk0 \dots sL} = E(R_{ijks_1}^* \dots R_{ijks_L}^*)$ be the correlation among the L components $R_{ijks_1}, \dots, R_{ijks_L}$ of R_{ijk} . For a given time point k , we employ the Bahadur representation (Bahadur, 1961; Cox, 1972) to express the joint probability

$$P(R_{ijk} = r_{ijk} | H_{ijk}^r, Y_i, X_i, Z_i) = \prod_{l=1}^{p_1} \left\{ \lambda_{ijkl}^{r_{ijkl}} (1 - \lambda_{ijkl})^{1-r_{ijkl}} \right\} \cdot \left\{ 1 + \sum_{1 \leq s < t \leq p_1} \rho_{ijkst} r_{ijks}^* r_{ijkt}^* + \dots + \rho_{ijk0 \dots p_1} r_{ijk0}^* \dots r_{ijkp_1}^* \right\}. \quad (17)$$

This strategy requires modeling the correlation structures of all orders. In practice, it is often the case that the second order dominates the association structure while the third and higher order association is null or negligible. Under such circumstances, we may conduct estimation along the lines of Sections 2 and 3. That is, for $l < l'$, let

$$\psi_{ijkl l'} = \frac{P(R_{ijkl} = 1, R_{ijkl'} = 1 | L_{ijk}^r) P(R_{ijkl} = 0, R_{ijkl'} = 0 | L_{ijk}^r)}{P(R_{ijkl} = 1, R_{ijkl'} = 0 | L_{ijk}^r) P(R_{ijkl} = 0, R_{ijkl'} = 1 | L_{ijk}^r)}$$

be the odds ratio featuring the association between R_{ijkl} and $R_{ijkl'}$, where

$L_{ijk}^r = \{H_{ijk}^r, Y_i, X_i, Z_i\}$. Regression models may be used to characterize the building blocks λ_{ijkl} and $\rho_{ijk l l'}$ (or equivalently, $\psi_{ijkl l'}$) in (17) where the third or higher order correlations are constrained to be zero.

For subjects j and j' in cluster i , we define $\rho_{i(jj')kst} = E(R_{ijks}^* R_{ij'kt}^*)$ as the conditional correlation between R_{ijks} and $R_{ij'kt}$. For given time k , we also employ the Bahadur representation (Bahadur, 1961; Cox, 1972) to express the joint probability

$$P(R_{ijk} = r_{ijk}, R_{ij'k} = r_{ij'k} | H_{ijk}^r, H_{ij'k}^r, Y_i, X_i, Z_i) = \prod_{l=1}^{p_1} \left\{ \lambda_{ijkl}^{r_{ijkl}} (1 - \lambda_{ijkl})^{1-r_{ijkl}} \lambda_{ij'kl}^{r_{ij'kl}} (1 - \lambda_{ij'kl})^{1-r_{ij'kl}} \right\} \cdot \left\{ 1 + \sum_{1 \leq s < t \leq p_1} \rho_{ijkst} r_{ijks}^* r_{ijkt}^* + \sum_{1 \leq s < t \leq p_1} \rho_{ij'kst} r_{ij'ks}^* r_{ij'kt}^* + \sum_{s=1}^{p_1} \sum_{t=1}^{p_1} \rho_{i(jj')kst} r_{ijks}^* r_{ij'kt}^* \right\}, \quad (18)$$

where we assume the third or higher correlations are zero. Similarly, we can also consider an odds ratio between R_{ijks} and $R_{ij'kt}$ to measure their association, and similar models can be constructed along the lines of Section 2.

For cluster i , the marginal probability π_{ijk} and the joint probability

$\pi_{ijk;j'k'} = P(R_{ijk}=1, R_{ij'k'}=1|Y_i, X_i, Z_i)$ can be expressed in terms of the conditional probabilities (17) and (18) as in Section 2.2.

Estimation of the response model parameters follows from the methods of Section 3.1, except that we replace $\Delta_f(\alpha)$ with

$$\Delta_i(\alpha) = \text{diag}\left(I(R_{ijk}=1) / \pi_{ijk}, \quad 1 \leq j \leq J_i, 1 \leq k \leq K\right),$$

and Δ_i^* with

$$\Delta_i^*(\alpha) = \text{diag}\left(I(R_{ijk}=1, R_{ij'k'}=1) / \pi_{ijk;j'k'}, (j, k) < (j', k')\right).$$

For estimation of the parameters of the missing data model, we construct two estimating equations in the same spirit as Section 3.2. Alternatively, one can solve an estimating equation similar to (7) with an independence correlation matrix, and then use the method of moments to obtain estimates for the correlations in (17) and (18). Though this method will lose some efficiency, it will reduce the computational burden considerably, especially when p_1 is large.

7. Discussion

With clustered longitudinal data, interests often lie in estimating the cross-sectional and longitudinal associations. In this paper, we consider analysis of clustered longitudinal data with missing covariates. We develop weighted first and second order estimating equations for data that are missing at random. To obtain consistent estimates of mean and association parameters, we build missing data models to accommodate cross-sectional associations as well as serial associations. Simulations demonstrate that ignoring these associations in the missing data models may induce biased estimates for the parameters in the response models.

In this paper, we focus the case that covariates are incompletely observed. Sometimes, the response variable and the covariates are missing. A comprehensive development on this problem has been given in Chen et al. (2010a) for longitudinal data with marginal mean regression model. This method can be extended to the case of clustered data as demonstrated in Section 6.

We have retained a focus on robust methods based on inverse probability weighted generalized estimating equations. Natural extensions include the development of doubly robust estimating equations (e.g., Van der Laan and Robins, 2003) or augmented estimating equations (e.g., Scharfstein et al., 1999). The former are of interest given the complex correlation structure needed in this setting where risk of misspecification is reasonably high. The latter are of interest given the considerable amount of data that may be discarded when individuals or visits may be discarded because any one of a number of covariates may be missing. Alternative strategy for dealing with incomplete data such as multiple imputation (e.g., Rubin, 1987; Schafer, 1997) may also warrant development in this setting.

Note that the missing at random assumption is not testable against the missing not at random model. With the missing not at random mechanism, estimating equation (7) and (8) cannot be generally used to estimate parameters associated with the missing data process due to the lack of information of unobserved measurements. However, our method can be readily

modified to perform sensitivity analyses to evaluate the impact of missingness on estimation of the response parameter τ . In particular, sensitivity analyses can be implemented using estimating equations (5) and (6) with the α parameters replaced by values that one wants to examine.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by Grants from the Natural Sciences and Engineering Research Council of Canada (G.Y. Yi and R.J. Cook) and the Canadian Institutes for Health Research (R.J. Cook). R.J. Cook is Canada Research Chair in Statistical Methods for Health Research. Dr. Xiao-Hua Zhou, Ph.D., is presently a Core Investigator and Biostatistics Unit Director at the Northwest HSR&D Center of Excellence, Department of Veterans Affairs Medical Center, Seattle, WA. Dr. Zhou's work was supported in part by U.S. Department of Veterans Affairs, Veterans Affairs Health Administration, HSR&D Grants (VA Research Career Scientist Award RCS 05-196), and the National Science Foundation of China (NSFC 30728019). Both Drs. Zhou and Chen were supported in part by National Institute on Aging Grant U01AG016976 to the National Alzheimer's Coordinating Center.

Appendix

Appendix A:

Supplementary data Supplementary data associated with this article can be found in the online version at <http://dx.doi.org.10.1016/j.jspi.2012.04.006>.

References

- Bahadur, RR. A representation of the joint distribution of responses to n dichotomous items. In: Solomon, H., editor. *Studies in Item Analysis and Prediction*, Stanford Mathematical Studies in the Social Sciences VI. Stanford University Press; Stanford, California: 1961. p. 158-168.
- Cox DR. The analysis of multivariate binary data. *Applied Statistics*. 1972; 21:113-120.
- Cameron R, Brown KS, Best JA, Pelkman CL, Madill CL, Manske SR, Payne ME. Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and social risk. *American Journal of Public Health*. 1999; 89:1827-1831. [PubMed: 10589310]
- Carey V, Zeger SL, Diggle PJ. Modeling multivariate binary data with alternating logistic regressions. *Biometrika*. 1993; 80:517-526.
- Chen B, Yi GY, Cook RJ. Weighted generalized estimating functions for incomplete longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association*. 2010a; 105:336-353.
- Chen B, Yi GY, Cook RJ. Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine*. 2010b; 29:1175-1189. [PubMed: 20437455]
- Chen B, Zhou XH. Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates. *Biometrics*. 2011; 67:830-842. [PubMed: 21281272]
- Cook RJ, Zeng L, Yi GY. Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*. 2004; 60:820-828. [PubMed: 15339307]
- Fitzmaurice GM, Laird NM, Rotnitzky A. Regression models for discrete longitudinal responses. *Statistical Science*. 1993; 8:284-299.
- Fitzmaurice GM, Lipsitz SR. A model for binary time series data with serial odds ratio patterns. *Applied Statistics*. 1995; 44:51-61.
- Fitzmaurice GM, Molenberghs G, Lipsitz SR. Regression models for longitudinal binary data responses with informative drop-outs. *Journal of the Royal Statistical Society, Series B*. 1995; 57:691-704.

- Lang J, Agresti AA. Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*. 1994; 89:625–632.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Liang KY, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal statistical Society, Series B*. 1992; 54:3–40.
- Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*. 1991; 78:153–160.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd ed. John Wiley and Sons, Inc.; 2002.
- Molenberghs G, Lesaffre. Marginal modeling of correlated ordinal data using an n-way Plackett distribution. *Journal of the American Statistical Association*. 1994; 89:633–644.
- Newey, WK.; Mcfadden, D. Large sample estimation and hypothesis testing. In: McFadden, D.; Engler, R., editors. *Handbook of Econometrics*. Vol. vol. 4. Elsevier; North-Holland, Amsterdam: 1994.
- Payment P, Richardson L, Siemiatycki J, Dewar R, Edwards M, Franco E. A randomized trial to evaluate the risk of gastrointestinal disease due to consumption of drinking water meeting current microbiological standards. *American Journal of Public Health*. 1991; 81:703–708. [PubMed: 2029037]
- Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics, Simulation and Computation*. 1994; 23:939–951.
- Perry CL, Kelder SH, Murray DM, Klepp KL. Community-wide smoking prevention: long-term outcomes of the Minnesota heart health program and the class of 1989 study. *American Journal of Public Health*. 1989; 82:1210–1216. [PubMed: 1503159]
- Prentice RL. Correlated binary regression with covariate specific to each binary observation. *Biometrics*. 1988; 44:1022–1048.
- Robins JM, Greenland S, Hu FC. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association*. 1999; 94:687–712.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressor are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*. 1995; 90:106–121.
- Rotnitzky A, Wypij D. A note on the bias of the estimators with missing data. *Biometrics*. 1994; 50:1163–1170. [PubMed: 7786998]
- Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc.; New York: 1987.
- Schafer, JL. *Analysis of Incomplete Multivariate Data*. Chapman and Hall; New York: 1997.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*. 1999; 94:1096–1120.
- Van der Laan, MJ.; Robins, JM. *Unified Methods for Censored Longitudinal Data and Causality*. Springer; New York: 2003.
- White H. Maximum likelihood estimation under misspecified models. *Econometrica*. 1982; 50:1–26.
- Yi GY, Cook RJ. Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*. 2002; 97:1071–1080.

Fig. 1.

Asymptotic percent relative bias of the mixed association parameter $\psi_{ij:kj'k'}(j, j', k, k')$, estimator with correctly specified marginal missing response model but where the cross-sectional association in the missing data processes are naively assumed to be conditionally independent.

Fig. 2.

Asymptotic percent relative bias of the cross-sectional association parameter $\psi_{i;jk;j^*k(j^*j)}$ estimator with correctly specified marginal missing response model but where the cross-sectional association in the missing data processes are naively assumed to be conditionally independent.

Table 1

Simulation results for the association study with missing covariates: about 20% missing ($\exp(\alpha_2) = 2.0, \exp(\alpha_3) = 1.5$)

Parameters	Independence weights				Clustered weights			
	Bias%	ASE	SD	CP%	Bias%	ASE	SD	CP%
$(\psi_{i,j2;j'2}^*, \psi_{i,j3;j'3}^*) = (4, 4)$								
β_0	0.74	0.106	0.105	94.3	1.84	0.105	0.105	94.8
β_1	0.14	0.114	0.114	94.6	0.72	0.114	0.114	95.4
β_2	0.73	0.124	0.125	95.2	0.42	0.125	0.125	94.2
β_3	0.25	0.140	0.138	94.4	-0.76	0.138	0.138	95.4
$\psi_{i,jk;j'k'}^*$	32.09	0.129	0.130	63.5	-0.51	0.126	0.127	95.4
$\psi_{i,jk;j'k}^*$	12.68	0.167	0.166	83.0	0.76	0.161	0.161	94.5
$\psi_{i,jk;jk'}^*$	-1.41	0.150	0.149	94.2	0.70	0.142	0.142	94.2
$(\psi_{i,j2;j'2}^*, \psi_{i,j3;j'3}^*) = (4, 2)$								
β_0	-2.71	0.111	0.111	94.3	-0.94	0.111	0.111	95.3
β_1	0.54	0.114	0.114	94.5	-0.14	0.115	0.114	94.9
β_2	0.49	0.127	0.126	94.4	0.24	0.126	0.126	94.6
β_3	0.65	0.127	0.128	95.5	0.98	0.128	0.128	94.8
$\psi_{i,jk;j'k'}^*$	31.63	0.132	0.132	63.8	-1.02	0.129	0.129	95.1
$\psi_{i,jk;j'k}^*$	11.02	0.171	0.171	83.4	0.08	0.167	0.167	94.6
$\psi_{i,jk;jk'}^*$	-2.33	0.156	0.154	94.6	0.06	0.147	0.146	94.4
$(\psi_{i,j2;j'2}^*, \psi_{i,j3;j'3}^*) = (2, 2)$								
β_0	-0.49	0.104	0.104	94.2	-0.24	0.104	0.104	95.1
β_1	0.86	0.114	0.114	94.8	-0.54	0.114	0.114	94.6
β_2	0.00	0.118	0.120	94.2	0.49	0.120	0.120	95.0
β_3	0.36	0.130	0.131	94.9	0.49	0.131	0.131	95.1
$\psi_{i,jk;j'k'}^*$	16.58	0.127	0.127	63.6	-1.24	0.125	0.125	94.5
$\psi_{i,jk;j'k}^*$	9.69	0.172	0.171	83.4	0.23	0.170	0.170	93.4
$\psi_{i,jk;jk'}^*$	-1.31	0.149	0.151	94.3	0.20	0.145	0.147	94.4
$(\psi_{i,j2;j'2}^*, \psi_{i,j3;j'3}^*) = (1, 1)$								
β_0	-2.19	0.106	0.106	94.4	-1.03	0.109	0.106	94.6
β_1	-0.86	0.112	0.111	94.4	0.79	0.110	0.111	94.5

Parameters	Independence weights				Clustered weights			
	Bias%	ASE	SD	CP%	Bias%	ASE	SD	CP%
β_2	2.71	0.126	0.126	94.4	0.73	0.124	0.126	94.5
β_3	1.23	0.129	0.131	94.3	1.02	0.131	0.131	94.6
$\psi_{i,jk;j'k'}$	-1.38	0.129	0.128	94.4	-1.09	0.128	0.128	94.4
$\psi_{i,jk;j'k}$	0.56	0.156	0.156	94.6	0.23	0.157	0.156	94.5
$\psi_{i,jk;jk'}$	-0.40	0.152	0.152	94.6	0.71	0.152	0.152	94.5

Table 2

Simulation results for the association study with missing covariates: about 35% missing ($\exp(\alpha_2) = 0.5, \exp(\alpha_3) = 1.0$)

Parameters	Independence weights				Clustered weights			
	Bias%	ASE	SD	CP%	Bias%	ASE	SD	CP%
$(\psi_{i,j2;j'2}^*, \psi_{i,j3;j'3}^*) = (4, 4)$								
β_0	-0.98	0.112	0.113	96.2	-0.96	0.113	0.113	95.4
β_1	-0.29	0.137	0.134	94.9	-0.31	0.135	0.134	94.9
β_2	0.73	0.130	0.130	93.4	0.79	0.130	0.130	94.8
β_3	-0.76	0.152	0.152	93.7	0.64	0.152	0.152	95.1
$\psi_{i,jk;j'k}$	52.97	0.234	0.234	53.4	1.38	0.213	0.214	95.2
$\psi_{i,jk;j'k}'$	18.34	0.320	0.317	82.9	-0.42	0.296	0.296	94.3
$\psi_{i,jk;jk}'$	2.041	0.309	0.308	94.4	-0.70	0.279	0.278	94.1
$(\psi_{i,j2;j'2}^*, \psi_{i,j3;j'3}^*) = (4, 2)$								
β_0	0.24	0.118	0.117	94.6	-0.98	0.117	0.117	95.3
β_1	-1.00	0.141	0.142	95.9	-0.54	0.141	0.142	95.5
β_2	1.07	0.128	0.127	94.1	0.91	0.127	0.127	93.9
β_3	-1.09	0.150	0.149	96.1	0.74	0.149	0.149	95.2
$\psi_{i,jk;j'k}'$	47.62	0.249	0.248	52.8	1.31	0.233	0.232	94.4
$\psi_{i,jk;j'k}$	23.18	0.295	0.293	82.6	-0.33	0.279	0.280	96.3
$\psi_{i,jk;jk}'$	1.76	0.304	0.305	94.3	-0.70	0.280	0.283	94.2
$(\psi_{i,j2;j'2}^*, \psi_{i,j3;j'3}^*) = (2, 2)$								
β_0	0.98	0.116	0.116	94.4	-1.00	0.116	0.116	94.4
β_1	-0.86	0.134	0.135	94.7	-0.43	0.135	0.135	93.2
β_2	0.12	0.124	0.124	94.5	0.37	0.124	0.124	94.1
β_3	-0.12	0.154	0.152	94.9	0.73	0.152	0.152	94.0
$\psi_{i,jk;j'k}'$	36.36	0.236	0.236	63.1	1.02	0.223	0.225	93.7
$\psi_{i,jk;j'k}$	13.27	0.287	0.287	83.3	-0.45	0.278	0.276	94.1
$\psi_{i,jk;jk}'$	1.62	0.297	0.296	93.9	-0.80	0.276	0.275	94.4
$(\psi_{i,j2;j'2}^*, \psi_{i,j3;j'3}^*) = (1, 1)$								
β_0	-1.08	0.118	0.117	94.6	-1.00	0.117	0.117	94.4
β_1	0.20	0.139	0.137	95.4	0.28	0.135	0.137	94.5

Parameters	Independence weights				Clustered weights			
	Bias%	ASE	SD	CP%	Bias%	ASE	SD	CP%
β_2	0.49	0.133	0.134	94.5	0.95	0.134	0.134	94.3
β_3	0.49	0.153	0.153	94.4	0.12	0.153	0.153	94.4
$\psi_{i,jk;j'k'}$	1.26	0.242	0.241	94.7	1.63	0.241	0.241	94.5
$\psi_{i,jk;j'k}$	-0.34	0.301	0.301	95.2	-0.45	0.301	0.301	94.6
$\psi_{i,jk;jk'}$	1.60	0.306	0.307	94.5	-0.80	0.307	0.307	94.5

Table 3

Results of estimation based on unweighted, independence weighted and clustered weighted GEE for the analysis of the National Alzheimer's Coordination Center Uniform Data Set: response model

Parameter	Unweighted			Independence weights			Clustered weights		
	Estimate	SE	<i>p</i> -Value	Estimate	SE	<i>p</i> -Value	Estimate	SE	<i>p</i> -Value
(Intercept)	0.3203	0.1604	0.0459	0.1550	0.1544	0.3155	0.1459	0.1543	0.3445
Sex(F)	-0.0230	0.0364	0.5266	-0.0146	0.0354	0.6798	-0.0145	0.0354	0.6810
CVCHF	0.0018	0.0909	0.9841	0.0207	0.0874	0.8125	0.0251	0.0875	0.7737
Depression	0.4040	0.0389	<0.0001	0.3997	0.0379	<0.0001	0.4009	0.0380	<0.0001
MMSE	-0.0209	0.0012	<0.0001	-0.0158	0.0010	<0.0001	-0.0159	0.0010	0.0000
FHDEM	-0.0078	0.0400	0.8454	-0.0071	0.0389	0.8535	-0.0069	0.0389	0.8596
Diabete	-0.1749	0.0533	0.0010	-0.1781	0.0516	0.0006	-0.1770	0.0516	0.0006
Hypert	-0.2183	0.0381	<0.0001	-0.2174	0.0371	<0.0001	-0.2188	0.0371	<0.0001
Educ	0.0021	0.0019	0.2844	0.0021	0.0018	0.2665	0.0021	0.0018	0.2582
Age	0.0124	0.0020	<0.0001	0.0130	0.0019	<0.0001	0.0131	0.0019	<0.0001
ϕ_0	0.0529	0.0048	<0.0001	0.0445	0.0046	<0.0001	0.0404	0.0045	<0.0001
ϕ^1	0.0726	0.0512	0.1565	0.0394	0.0467	0.3991	0.0294	0.0465	0.5271
ϕ_2	0.1920	0.0156	<0.0001	0.2097	0.0148	<0.0001	0.2140	0.0144	<0.0001

Table 4

Results of estimation based on independence weighted and clustered weighted GEE for the analysis of the National Alzheimer's Coordination Center Uniform Data Set: missing data models

Parameter	Independence weights			Clustered weights		
	Estimate	SE	<i>p</i> -Value	Estimate	SE	<i>p</i> -Value
(Intercept)	5.5915	0.3740	<0.0001	5.3232	0.3959	<0.0001
Sex(F)	−0.2558	0.0796	0.0013	−0.2553	0.0797	0.0014
CVCHF	−0.2439	0.1622	0.1327	−0.2259	0.1634	0.1666
Prev. Depre	19.8033	6.6579	0.0029	19.7987	6.8410	0.0034
MMSE	−0.0244	0.0015	<0.0001	−0.0170	0.0043	0.0001
FHDEM	0.0540	0.0893	0.5452	0.0586	0.0894	0.5121
Diabete	−0.2348	0.1117	0.0355	−0.2342	0.1118	0.0363
Hypert	0.0763	0.0822	0.3532	0.0726	0.0823	0.3774
Educ	−0.0025	0.0035	0.4887	−0.0023	0.0035	0.5224
Age	−0.0377	0.0044	<0.0001	−0.0371	0.0044	<0.0001
Vist2	0.8691	0.1188	<0.0001	1.2009	0.1871	<0.0001
Vist3	1.3170	0.1271	<0.0001	1.5359	0.1993	<0.0001
Vist4	1.6206	0.1547	<0.0001	1.6402	0.2460	<0.0001
ϕ_2^*				0.4628	0.0037	<0.0001
ϕ_3^*				0.3764	0.0068	<0.0001
ϕ_4^*				0.4089	0.0026	<0.0001